Similarity Measurement of XML Documents Based on Structure and Contents

Tae-Soon Kim¹, Ju-Hong Lee^{1,*}, Jae-Won Song¹, and Deok-Hwan Kim²

¹Dept. of Computer Science & Information Engineering, Inha University, Incheon, Korea

²Dept. of Electronics Engineering, Inha University
{kts429, sjw}@datamining.inha.ac.kr,
{juhong, deokhwan}@inha.ac.kr

Abstract. Researches on the similarity measure between XML documents are being progressed in order to effectively control and retrieve various XML documents. Previous works mostly suggest similarity-measuring methods focusing only on the tag structure of XML documents. However, they have a problem of incorrectly calculating the semantic similarity of XML contents. In this paper, we propose a new similarity measurement method considering not only the structural information of tags in XML documents but also the semantic information of tags and text content information related with the tags. Our experiments demonstrate that our proposed method improves the accuracy of similarity, compared to the previous works.

1 Introduction

As the use of XML [2] is increased, many researches for calculating and controlling the similarity between XML documents are being actively progressed in order for efficient document organization, navigation and retrieval of a large amount of documents [1,5,6,9,12].

Lian et al. [5] summarized XML documents into an S-graph which reflects structural information. It is used to measure the similarity using only the tag structural information of XML documents. Patrick et al. [9] proposed a measure on the structural similarity among XML documents and DTDs, which is natural to understand and fast to calculate. Bergamashi et al. [1] proposed a method of calculating the similarity between DTDs or schemas of XML documents. These previous works presented the similarity measure using only the structural information of tags. Our preliminary work [3] proposed the vector model to reflect the structural information and semantic information of tags in XML documents.

In this paper, in order to increase the accuracy of semantic similarity between XML documents, we suggest the enhanced similarity measurement method using vector model which considers not only the structural information of tags but also the semantic information of tags and the text content information of XML documents

* Corresponding author.

since the semantic information of XML documents are included in tag structure and text content within tags.

This paper is organized as follows: Chapter 2 describes the vector model combining path term and content term and presents a new similarity measure for XML documents. Chapter 3 shows various experiments and result analysis. Finally, Chapter 4 concludes this work.

2 Vector Model for Similarity Between XML Documents

Our preliminary work proposed the vector model which considers the tag path as a term to reflect the structural information and semantic information of tags in XML documents simultaneously [3]. But, text contents of an XML document, differently from a general document, is included in a tag. Since the words in the text with the same frequency may have different importances depending on the tag in which they are, the weight of the word in the text must be reflected appropriately according to the importance of the related tags. As an example, in the XML document including some research paper, some tags such as title and abstract have relatively a few words, but the importance of the words within this tags are high. Some tags such as Introduction may have more words than others, but the importances of their words are relatively lower than those of words within *title* and *abstract* tags.

We call the text words within tags as **content terms**. *CTW* (Content Term Weight) is a weight value of a content term in a XML document. The *CTW* is the product of *CTWF* (Content Term Weighted Frequency) and *CTIDF* (Content Term Inverse Document Frequency). CTW_{ij} of j^{th} content term in i^{th} document is shown in equation (1).

$$CTW_{ij} = CTWF_{ij} \times CTIDF_j \quad (1)$$

$CTWF_{ij}$ is shown in equation (2).

$$CTWF_{ij} = \sum_{k=1}^n \left(cfreq_{ikj} \times \left(1 + \log \left(\frac{\# \text{content term in } i^{\text{th}} \text{ document}}{\# \text{content term in } k^{\text{th}} \text{ tag}} \right) \right) \right) \quad (2)$$

Where k is the index of tags in a document and $cfreq_{ikj}$ is the frequency of j^{th} content term included in k^{th} tag of i^{th} document. $(1 + \log(\# \text{content term in } i^{\text{th}} \text{ document} / \# \text{content term in } k^{\text{th}} \text{ tag}))$ is the formula to calculate the tag importance based on the number of included words. Note that the log function is used to lower the differences of values.

$CTIDF_j$ is shown in equation (3)

$$CTIDF_j = \log \frac{N}{DF_j} \quad (3)$$

Where N is the total number of documents, and DF_j is the number of documents in which the j^{th} content term appears.

The similarity using *CTW* is called **content term similarity (CTS)**. In vector model, the similarity measure between document d_i and document d_j can be quantified as the cosine value of adjacent angle between two vectors as shown in equation (4).

$$CTS(d_i, d_j) = \frac{D_i \cdot D_j}{|D_i| \times |D_j|} = \frac{\sum_{k=1}^t CTW_{ki} \times CTW_{kj}}{\sqrt{\sum_{k=1}^t CTW_{ki}^2} \times \sqrt{\sum_{k=1}^t CTW_{kj}^2}} \quad (4)$$

where t refers to the number of total terms within d_i and d_j , $D_i = (CTW_{1i}, CTW_{2i}, \dots, CTW_{ti})$, and $D_j = (CTW_{1j}, CTW_{2j}, \dots, CTW_{tj})$, respectively [10].

The similarity measure of the general vector model considers text content only. In the case of the XML document, however, we should consider the importance of tags to measure the similarity between XML documents. The overall similarity between XML documents is calculated as the weighted sum of path term similarity (*PTS*) and content term similarity (*CTS*) shown in equation (5). Here, the weights of *PTS* and the weights of *CTS* are represented as pw and cw , respectively.

$$\begin{cases} sim(d_1, d_2) = pw \times PTS(d_1, d_2) + cw \times CTS(d_1, d_2) \\ pw + cw = 1 \end{cases} \quad (5)$$

3 Test and Evaluation

In experiments, we used 1200 XML documents taken from SIGMOD [11], O'REILLY [8], and NIAGARA [7]. We applied the K-means algorithm using the similarity of general vector model, the similarity of S-graph based method, and the similarity of the proposed method. Table. 1 illustrates their purity values [4]. The experimental results demonstrate that the proposed method yields more accurate performance than S-graph based method and the general vector model.

Table 1. Performance comparison of various methods in terms of purity

	similarity of general vector model	S-graph	Proposed method
SIGMOD	0.4728	0.5055	0.6678
NIAGARA	0.5385	0.8154	0.8912
O'REILLY	0.4682	0.7533	0.8723

4 Conclusion

The previous methods are mostly based on graph since they considered only structural information of tags in XML documents. In order to measure more accurate semantic similarity, the semantic information of tags, and the text content information must be included in the similarity measure as well as the structural information of tags. In this paper, therefore, we introduce a new effective semantic similarity measure between

XML documents, which uses the vector model in order to reflect three kinds of information in the similarity measure simultaneously. This method undoubtedly can measure more accurate semantic similarity between XML documents because almost all information about semantic meaning of XML documents is reflected in the similarity measure.

Acknowledgement

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment).

References

1. Bergamashi, S., Castano, S.: Semantic Integration of Semistructured and Structured Data Sources. *SIGMOD Record* **28** (1999) pp. 54-59.
2. Brau, B., Bray, J.: Extensible Markup Language (XML) 1.0. W3C Recommendation, <http://www.w3.org/Tr/Rec-xml>
3. Kim, T.S., Lee, J.H.: Semantic Structural Similarity for Clustering XML documents. Inha University Technical Report (2006), http://webbase.inha.ac.kr/TechnicalReport/tech_04.pdf
4. Li, T., Ma, S.: Document Clustering via Adaptive Subspace Iteration. the 27th annual international ACM SIGIR conference on Research and development in information retrieval, Sheffield, United Kingdom (2004) pp. 218-225.
5. Lian, W. , Wai-lok, D.: An Efficient and Scalable Algorithm for Clustering XML Documents by Structure. *IEEE Transactions on Knowledge and Data Engineering* **16** (2004) pp. 82-96.
6. Marcos, C., Lozano, S.: A DTD for an XML-Based Mathematical Modeling Language. *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg **2660** (2003) pp. 968-977.
7. NIAGARA Experimental Data, <http://www.cs.wisc.edu/niagara/data.html>
8. O'REILLY XML Data, http://www.xml.com/pub/rg/Example_Files
9. Patrick, K.L., Ng, Vincent T.: Structural Similarity between XML Documents and DTDs. *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg **2659** (2003) pp. 412-241.
10. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. *Information processing and Management: an International journal* **24** (1988) pp.513-524
11. *SIGMOD Record* data, <http://www.acm.org/sigs/sigmod/record/xml>
12. Young, K.N., Joseph G., Guilian W.: A Metadata Tool for Retrieval from Heterogeneous Distributed XML Documents. *Lecture Notes in Computer Science*, Springer-Verlag, Berlin Heidelberg **2660** (2003) pp. 1020-1029.